Regret Bounds for Restless Markov Bandits

Ronald Ortner[1,2], Daniil Ryabko[2], Peter Auer[1], and Rémi Munos[2]

[1] Montanuniversitaet Leoben
[2] INRIA Lille-Nord Europe, équipe SequeL
{rortner,auer}@unileoben.ac.at, daniil@ryabko.net, remi.munos@inria.fr

Abstract. We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm that after T steps achieves $\tilde{O}(\sqrt{T})$ regret with respect to the best policy that knows the distributions of all arms. No assumptions on the Markov chains are made except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

1 Introduction

In the bandit problem the learner has to decide at time steps $t = 1, 2, \ldots$ which of the finitely many available arms to pull. Each arm produces a reward in a stochastic manner. The goal is to maximize the reward accumulated over time.

Following [1], traditionally it is assumed that the rewards produced by each given arm are independent and identically distributed (i.i.d.). If the probability distributions of the rewards of each arm are known, the best strategy is to only pull the arm with the highest expected reward. Thus, in the i.i.d. bandit setting the *regret* is measured with respect to the best arm. An extension of this setting is to assume that the rewards generated by each arm are not i.i.d., but are governed by some more complex stochastic process. Markov chains suggest themselves as an interesting and non-trivial model. In this setting it is often natural to assume that the stochastic process (Markov chain) governing each arm does not depend on the actions of the learner. That is, the chain takes transitions independently of whether the learner pulls that arm or not (giving the name *restless bandit* to the problem). The latter property makes the problem rather challenging: since we are not observing the state of each arm, the problem becomes a partially observable Markov decision process (POMDP), rather than being a (special case of) a fully observable MDP, as in the traditional i.i.d. setting. One of the applications that motivate the restless bandit problem is the so-called *cognitive radio* problem (e.g., [2]): Each arm of the bandit is a radio channel that can be busy or available. The learner (an appliance) can only sense a certain number of channels (in the basic case only a single one) at a time, which is equivalent to pulling an arm. It is natural to assume that whether the channel is busy or not at a given time step depends on the past — so a Markov chain is the simplest realistic model — but does not depend on which channel the appliance is sensing. (See also Example 1 in Section 3 for an illustration of a simple instance of this problem.)

What makes the restless Markov bandit problem particularly interesting is that *one can do much better than pulling the best arm*. This can be seen already on simple examples with two-state Markov chains (see Section 3 below). Remarkably, this feature is often overlooked, notably by some early work on restless bandits, e.g. [3], where the regret is measured with respect to the mean reward of the best arm. This feature also makes the problem more difficult and in some sense more general than the non-stochastic bandit problem, in which the regret usually is measured with respect to the best arm in hindsight [4]. Finally, it is also this feature that makes the problem principally different from the so-called *rested* bandit problem, in which each Markov chain only takes transitions when the corresponding arm is pulled.

Thus, in the restless Markov bandit problem that we study, the regret should be measured not with respect to the best arm, but with respect to the best policy knowing the distribution of all arms. To understand what kind of regret bounds can be obtained in this setting, it is useful to compare it to the i.i.d. bandit problem and to the problem of learning an MDP. In the i.i.d. bandit problem, the minimax regret expressed in terms of the horizon T and the number of arms only is $O(\sqrt{T})$, cf. [5]. If we allow problem-dependent constants into consideration, then the regret becomes of order $\log T$ but depends also on the gap between the expected reward of the best and the second-best arm. In the problem of learning to behave optimally in an MDP, nontrivial problem-independent finite-time regret guarantees (that is, regret depending only on T and the number of states and actions) are not possible to achieve. It is possible to obtain $O(\sqrt{T})$ regret bounds that also depend on the diameter of the MDP [6] or similar related constants, such as the span of the optimal bias vector [7]. Regret bounds of order $\log T$ are only possible if one additionally allows into consideration constants expressed in terms of policies, such as the gap between the average reward obtained by the best and the second-best policy [6]. The difference between these constants and constants such as the diameter of an MDP is that one can try to estimate the latter, while estimating the former is at least as difficult as solving the original problem — finding the best policy. Turning to our restless Markov bandit problem, so far, to the best of our knowledge no regret bounds are available for the general problem. However, several special cases have been considered. Specifically, $O(\log T)$ bounds have been obtained in [8] and [9]. While the latter considers the two-armed restless bandit case, the results of [8] are constrained by some ad hoc assumptions on the transition probabilities and on the structure of the optimal policy of the problem. Also the dependence of the regret bound on the problem parameters is unclear, while computational aspects of the algorithm (which alternates exploration and exploitation steps) are neglected. Finally, while regret bounds for the Exp3.S algorithm [4] could be applied, these depend on the "hardness" of the reward sequences, which in the case of reward sequences generated by a Markov chain can be arbitrarily high.

Here we present an algorithm for which we derive $\tilde{O}(\sqrt{T})$ regret bounds, making no assumptions on the distribution of the Markov chains. The algorithm is based on constructing an approximate MDP representation of the POMDP problem, and then using a modification of the UCRL2 algorithm of [6] to learn this approximate MDP. In addition to the horizon T and the number of arms and states, the regret bound also depends on the diameter and the mixing time (which can be eliminated however) of the Markov chains of the arms. If the regret has to be expressed only in these terms, then our lower bound shows that the dependence on T cannot be significantly improved.

2 Preliminaries

Given are K arms, where underlying each arm j there is an irreducible Markov chain with state space S_j and transition matrix P_j. For each state s in S_j there are mean rewards $r_j(s)$, which we assume to be bounded in $[0,1]$. For the time being, we will assume that the learner knows the number of states for each arm and that all Markov chains are aperiodic. In Section 7, we discuss periodic chains, while in Section 8 we indicate how to deal with unknown state spaces. In any case, the learner knows neither the transition probabilities nor the mean rewards.

For each time step $t = 1, 2, \ldots$ the learner chooses one of the arms, observes the current state s of the chosen arm i and receives a random reward with mean $r_i(s)$. After this, the state of each arm j changes according to the transition matrices P_j. The learner however is not able to observe the current state of the individual arms. We are interested in competing with the optimal policy π^* which knows the mean rewards and transition matrices, yet observes as the learner only the current state of the chosen arm. Thus, we are looking for algorithms which after any T steps have small regret with respect to π^*, i.e. minimize

$$T \cdot \rho^* - \textstyle\sum_{t=1}^{T} r_t,$$

where r_t denotes the (random) reward earned at step t and ρ^* is the average reward of the optimal policy π^*. (It will be seen in Section 5 that π^* and ρ^* are indeed well-defined.)

Mixing Times and Diameter If an arm j is not selected for a large number of time steps, the distribution over states when selecting j will be close to the stationary distribution μ_j of the Markov chain underlying arm j. Let μ_s^t be the distribution after t steps when starting in state $s \in S_j$. Then setting

$$d_j(t) := \max_{s \in S_j} \|\mu_s^t - \mu_j\|_1 := \max_{s \in S_j} \sum_{s' \in S_j} |\mu_s^t(s') - \mu_j(s')|,$$

we define the ε-*mixing time* of the Markov chain as

$$T_{\text{mix}}^j(\varepsilon) := \min\{t \in \mathbb{N} \mid d_j(t) \leq \varepsilon\}.$$

Setting somewhat arbitrarily *the* mixing time of the chain to $T^j_{\mathrm{mix}} := T^j_{\mathrm{mix}}(\frac{1}{4})$, one can show (cf. eq. 4.36 in [10]) that

$$T^j_{\mathrm{mix}}(\varepsilon) \leq \lceil \log_2 \tfrac{1}{\varepsilon} \rceil \cdot T^j_{\mathrm{mix}}. \tag{1}$$

Finally, let $T_j(s, s')$ be the expected time it takes in arm j to reach s' when starting in s. We set the *diameter* of arm j to be $D_j := \max_{s,s' \in S_j} T_j(s, s')$.

3 Examples

Next we present a few examples that give insight into the nature of the problem and the difficulties in finding solutions. In particular, the examples demonstrate that (i) the optimal reward can be (much) bigger than the average reward of the best arm, (ii) the optimal policy does not maximize the immediate reward, (iii) the optimal policy cannot always be expressed in terms of arm indexes.

Example 1. In this example the average reward of each of the two arms of a bandit is $\frac{1}{2}$, but the reward of the optimal policy is close to $\frac{3}{4}$. Consider a two-armed bandit. Each arm has two possible states, 0 and 1, which are also the rewards. Underlying each of the two arms is a (two-state) Markov chain with transition matrix $\begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$, where ϵ is small. Thus, a typical trajectory of each arm looks like this: 000000000001111111111111111000000000..., and the average reward for each arm is $\frac{1}{2}$. It is easy to see that the optimal policy starts with any arm, and then switches the arm whenever the reward is 0, and otherwise sticks to the same arm. The average reward is close to $\frac{3}{4}$ — much larger than the reward of each arm.

This example has a natural interpretation in terms of *cognitive radio*: two radio channels are available, each of which can be either busy (0) or available (1). A device can only sense (and use) one channel at a time, and one wants to maximize the amount of time the channel it tries to use is available.

Example 2. Consider the previous example, but with ϵ close to 1. Thus, a typical trajectory of each arm is now 01010101001010110..., and the optimal policy switches arms if the previous reward was 1 and stays otherwise.

Example 3. In this example the optimal policy does not maximize the immediate reward. Again, consider a two-armed bandit. Arm 1 is as in Example 1, and arm 2 provides Bernoulli i.i.d. rewards with probability $\frac{1}{2}$ of getting reward 1. The optimal policy (which knows the distributions) will sample arm 1 until it obtains reward 0, when it switches to arm 2. However, it will sample arm 1 again after some time t (depending on ϵ), and only switch back to arm 2 when the reward on arm 1 is 0. Note that whatever t is, the expected reward for choosing arm 1 will be strictly smaller than $\frac{1}{2}$, since the last observed reward was 0 and the limiting probability of observing reward 1 (when $t \to \infty$) is $\frac{1}{2}$. At the same time, the expected reward of the second arm is always $\frac{1}{2}$. Thus, the optimal policy will sometimes "explore" by pulling the arm with the smaller expected reward.

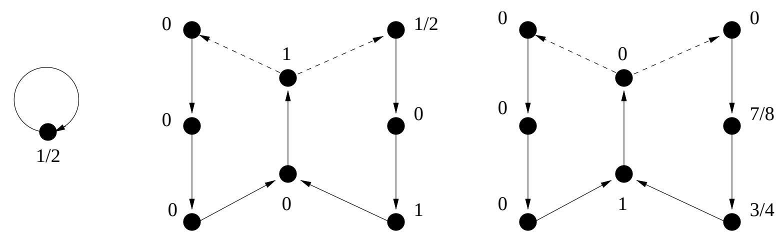

Fig. 1. *Example 4.* Dashed transitions are with probability $\frac{1}{2}$, others are deterministic with probability 1. Numbers are rewards in the respective state.

An intuitively appealing idea is to look for an optimal policy in an *index* form. That is, for each arm the policy maintains an index which is a function of time, states, and rewards *of this arm only.* At each time step, the policy samples the arm that has maximal index. This seems promising for at least two reasons: First, the distributions of the arms are assumed independent, so it may seem reasonable to evaluate them independently as well; second, this works in the i.i.d. case (e.g., the Gittins index [11] or UCB [12]). This idea also motivates the setting when just one out of two arms is Markov and the other is i.i.d., see e.g. [9]. Index policies for restless Markov bandits were also studied in [13]. Despite their intuitive appeal, in general, index policies are suboptimal.

Theorem 1. *For each index-based policy π there is a restless Markov bandit problem in which π behaves suboptimally.*

Proof. Consider the three bandits L (left), C (center), and R (right) in Figure 1, where C and R start in the 1 reward state. (Arms C and R can easily be made aperiodic by adding further sufficiently small transition probabilities.) Assume that C has been observed in the $\frac{1}{2}$ reward state one step before, while R has been observed in the 1 reward state three steps ago. The optimal policy will choose arm L which gives reward $\frac{1}{2}$ with certainty (C gives reward 0 with certainty, while R gives reward $\frac{7}{8}$ with probability $\frac{1}{2}$) and subsequently arms C and R. However, if arm C was missing, in the same situation, the optimal policy would choose R: Although the immediate expected reward is smaller than when choosing L, sampling R gives also information about the current state, which can earn reward $\frac{3}{4}$ a step later. Clearly, no index based policy will behave optimally in both settings. □

4 Main Results

Theorem 2. *Consider a restless bandit with K aperiodic arms having state spaces S_j, diameters D_j, and mixing times T^j_{mix} ($j = 1, \ldots, K$). Then with probability at least $1-\delta$ the regret of Algorithm 2 (presented in Section 5 below) after T steps is upper bounded by*

$$\text{const} \cdot S \cdot T_{\text{mix}}^{3/2} \cdot \textstyle\prod_{j=1}^{K}(4D_j) \cdot \max_i \log(D_i) \cdot \log^2\left(\frac{T}{\delta}\right) \cdot \sqrt{T},$$

where $S := \sum_{j=1}^{K} |S_j|$ *is the total number of states and* $T_{\mathrm{mix}} := \max_j T_{\mathrm{mix}}^j$ *the maximal mixing time. Further, the dependence on* T_{mix} *can be eliminated to show that with probability at least* $1-\delta$ *the regret is bounded by*

$$O\left(S \cdot \prod_{j=1}^{K}(4D_j) \cdot \max_i \log(D_i) \cdot \log^{7/2}\left(\tfrac{T}{\delta}\right) \cdot \sqrt{T}\right).$$

Remark 1. For periodic chains the bound of Theorem 2 has worse dependence on the state space, for details see Remark 5 in Section 7.

Theorem 3. *For any algorithm, any* $K > 1$ *and any* $m \geq 1$ *there is a* K*-armed restless bandit problem with a total number of* $S := Km$ *states, such that the regret after* T *steps is lower bounded by* $\Omega(\sqrt{ST})$.

Remark 2. While it is easy to see that lower bounds depend on the total number of states over all arms, the dependence on other parameters in our upper bound is not clear. For example, intuitively, while in the general MDP case one wrong step may cost up to D — the MDP's diameter [6] — steps to compensate for, here the Markov chains evolve independently of the learner's actions, and the upper bound's dependence on the diameter may be just an artefact of the proof.

5 Constructing the Algorithm

MDP Representation We represent the setting as an MDP by recalling for each arm the last observed state and the number of time steps which have gone by since this last observation. Thus, each state of the MDP representation is of the form $(s_j, n_j)_{j=1}^K := (s_1, n_1, s_2, n_2, \ldots, s_K, n_K)$ with $s_j \in S_j$ and $n_j \in \mathbb{N}$, meaning that each arm j has not been chosen for n_j steps when it was in state s_j. More precisely, $(s_j, n_j)_{j=1}^K$ is a state of the considered MDP if and only if (i) all n_j are distinct and (ii) there is a j with $n_j = 1$.[3] The action space of the MDP is $\{1, 2, \ldots, K\}$, and the transition probabilities from a state $(s_j, n_j)_{j=1}^K$ are given by the n_j-step transition probabilities $p_j^{(n_j)}(s, s')$ of the Markov chain underlying the chosen arm j (these are defined by the matrix power of the single step transition probability matrix, i.e. $P_j^{n_j}$). That is, the probability for a transition from state $(s_j, n_j)_{j=1}^K$ to $(s'_j, n'_j)_{j=1}^K$ under action j is given by $p_j^{(n_j)}(s_j, s'_j)$ iff (i) $n'_j = 1$, (ii) $n'_\ell = n_\ell + 1$ and $s_\ell = s'_\ell$ for all $\ell \neq j$. All other transition probabilities are 0. Finally, the mean reward for choosing arm j in state $(s_j, n_j)_{j=1}^K$ is given by $\sum_{s \in S_j} p_j^{(n_j)}(s_j, s) \cdot r_j(s)$. This MDP representation has already been considered in [8].

Obviously, within T steps any policy can reach only states with $n_j \leq T$. Correspondingly, if we are interested in the regret within T steps, it will be sufficient to consider the finite sub-MDP consisting of states with $n_j \leq T$. We call this the T*-step representation* of the problem, and the regret will be measured with respect to the optimal policy in this T-step representation.

[3] Actually, one would need to add for each arm j with $|S_j| > 1$ a special state for not having sampled j so far. However, for the sake of simplicity we assume that in the beginning each arm is sampled once. The respective regret is negligible.

Algorithm 1 The colored UCRL2 algorithm

Input: Confidence parameter $\delta > 0$, aggregation parameter $\varepsilon > 0$, state space S, action space A, coloring and translation functions, a bound B on the size of the support of transition probability distributions.

Initialization: Set $t := 1$, and observe the initial state s_1.

for episodes $k = 1, 2, \ldots$ **do**

Initialize episode k:

Set the start time of episode k, $t_k := t$. Let $N_k(c)$ be the number of times a state-action pair of color c has been visited prior to episode k, and $v_k(c)$ the number of times a state-action pair of color c has been visited in episode k. Compute estimates $\hat{r}_k(s,a)$ and $\hat{p}_k(s'|s,a)$ for rewards and transition probabilities, using all samples from state-action pairs of the same color $c(s,a)$, respectively.

Compute policy $\tilde{\pi}_k$:

Let $\mathcal{M}_k$ be the set of plausible MDPs with rewards $\tilde{r}(s,a)$ and transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying

$$\left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| \;\leq\; \varepsilon + \sqrt{\tfrac{7\log(2Ct_k/\delta)}{2\max\{1, N_k(c(s,a))\}}}, \tag{2}$$

$$\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \;\leq\; \varepsilon + \sqrt{\tfrac{56B\log(4Ct_k/\delta)}{\max\{1, N_k(c(s,a))\}}}\,, \tag{3}$$

where C is the number of distinct colors. Let $\rho(\pi, M)$ be the average reward of a policy $\pi : S \to A$ on an MDP $M \in \mathcal{M}_k$. Choose (e.g. by extended value iteration [6]) an optimal policy $\tilde{\pi}_k$ and an optimistic $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\rho(\tilde{\pi}_k, \tilde{M}_k) = \max\{\rho(\pi, M) \,|\, \pi : S \to A,\, M \in \mathcal{M}_k\}. \tag{4}$$

Execute policy $\tilde{\pi}_k$:

while $v_k(c(s_t, \tilde{\pi}_k(s_t))) < \max\{1, N_k(c(s_t, \tilde{\pi}_k(s_t)))\}$ **do**

▷ Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t, and observe next state s_{t+1}.

▷ Set $t := t + 1$.

end while

end for

Structure of the MDP Representation The MDP representation of our problem has some special structural properties. In particular, rewards and transition probabilities for choosing arm j only depend on the state of arm j, i.e. s_j and n_j. Moreover, the support for each transition probability distribution is bounded, and for $n_j \geq T^j_{\text{mix}}(\varepsilon)$ the transition probability distribution will be close to the stationary distribution of arm j. Thus, one could reduce the T-step representation further by aggregating states[4] $(s_j, n_j)_{j=1}^K$, $(s'_j, n'_j)_{j=1}^K$ whenever $n_j, n'_j \geq T^j_{\text{mix}}(\varepsilon)$ and $s_\ell = s'_\ell$, $n_\ell = n'_\ell$ for $\ell \neq j$. The rewards and transition probability distributions of aggregated states are ε-close, so that the error by

[4] Aggregation of states $s_1, \ldots, s_n$ means that these states are replaced by a new state s_{agg} inheriting rewards and transition probabilities from an arbitrary s_i (or averaging over all s_j). Transitions to this state are set to $p(s_{\text{agg}}|s,a) := \sum_j p(s_j|s,a)$.

Algorithm 2 The restless bandits algorithm

Input: Confidence parameter $\delta > 0$, the number of states S_j and mixing time T^j_{mix} of each arm j, horizon T.

$\triangleright$ Choose $\varepsilon = 1/\sqrt{T}$ and execute colored UCRL2 (with confidence parameter δ) on the ε-structured MDP described in the "coloring" paragraph at the end of Section 5.

aggregation can be bounded by results given in [14]. While this is helpful for approximating the problem when all parameters are known, it cannot be used directly when learning, since the observations in the aggregated states do not correspond to an MDP anymore. Thus, while standard reinforcement learning algorithms are still applicable, there are no theoretical guarantees for them.

ε-structured MDPs and Colored UCRL2 In the following, we exploit the special structure of the MDP representation. We generalize some of its structural properties in the following definition.

Definition 1. *An ε-structured MDP is an MDP with finite state space S, finite action space A, transition probability distributions $p(\cdot|s,a)$, mean rewards $r(s,a) \in [0,1]$, and a coloring function $c : S \times A \to \mathcal{C}$, where $\mathcal{C}$ is a set of colors. Further, for each two pairs (s,a), $(s',a') \in S \times A$ with $c(s,a) = c(s',a')$ there is a bijective translation function $\phi_{s,a,s',a'} : S \to S$ such that $\sum_{s''} \left|p(s''|s,a) - p(\phi_{s,a,s',a'}(s'')|s',a')\right| < \varepsilon$ and $|r(s,a) - r(s',a')| < \varepsilon$.*

If there are states s, s' in an ε-structured MDP such that $c(s,a) = c(s',a)$ for all actions a and the associated translation function $\phi_{s,a,s',a}$ is the identity, we may aggregate the states (cf. footnote 4). We call the MDP in which all such states are aggregated the *aggregated ε-structured MDP*.

For learning in ε-structured MDPs we consider a modification of the UCRL2 algorithm of [6]. The *colored* UCRL2 algorithm is shown in Figure 1. As the original UCRL2 algorithm it maintains confidence intervals for rewards and transition probabilities which define a set of plausible MDPs $\mathcal{M}_k$. In each episode k, the algorithm chooses an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ and an optimal policy which maximize the average reward, cf. (4). Colored UCRL2 calculates estimates from all samples of state-action pairs of the same color, and works with respectively adapted confidence intervals and a corresponding adapted episode termination criterion. Basically, an episode ends when for some color c the number of visits in state-action pairs of color c has doubled.

Coloring the T-step representation Now, we can turn the T-step representation into an ε-structured MDP, assigning the same color to state-action pairs where the chosen arm is in the same state, that is, $c((s_i,n_i)_{i=1}^K, j) = c((s'_i,n'_i)_{i=1}^K, j')$ iff $j = j'$, $s_j = s'_j$, and either $n_j = n'_j$ or $n_j, n'_j \geq T^j_{\text{mix}}(\varepsilon)$. The translation functions are chosen accordingly. This ε-structured MDP can be learned with colored UCRL2, see Algorithm 2, our restless bandits algorithm.

(The dependence on the horizon T and the mixing times T^j_{mix} as input parameters can be eliminated, cf. the proof of Theorem 2 in Section 7.)

6 Regret Bounds for Colored UCRL2

The following is a generalization of the regret bounds for UCRL2 to ε-structured MDPs. The theorem gives improved (with respect to UCRL2) bounds if there are only a few parameters to estimate in the MDP to learn. Recall that the *diameter* of an MDP is the maximal expected transition time between any two states (choosing an appropriate policy), cf. [6].

Theorem 4. *Let M be an ε-structured MDP with finite state space S, finite action space A, transition probability distributions $p(\cdot|s,a)$, mean rewards $r(s,a) \in [0,1]$, coloring function c and associate translation functions. Assume the learner has complete knowledge of state-action pairs $\Psi_K \subseteq S \times A$, while the state-action pairs in $\Psi_U := S \times A \setminus \Psi_K$ are unknown and have to be learned. However, the learner knows c and all associate translation functions as well as an upper bound B on the size of the support of each transition probability distribution in Ψ_U. Then with probability at least $1-\delta$, after any T steps colored* UCRL2 [5] *gives regret upper bounded by*

$$42 D_\varepsilon \sqrt{B C_U T \log\left(\tfrac{T}{\delta}\right)} + \varepsilon(D_\varepsilon + 2)T,$$

where C_U is the total number of colors for states in Ψ_U, and D_ε is the diameter of the aggregated ε-structured MDP.

The proof of this theorem is given in the appendix.

Remark 3. For $\varepsilon = 0$, one can also obtain logarithmic bounds analogously to Theorem 4 of [6]. With no additional information for the learner one gets the original UCRL2 bounds (with a slightly larger constant), trivially choosing B to be the number of states and assigning each state-action pair an individual color.

7 Proofs

We start with bounding the diameter in the aggregated ε-structured MDP.

Lemma 1. *For $\varepsilon \leq 1/4$, the diameter D_ε in the aggregated ε-structured MDP can be upper bounded by $2\left\lceil \log_2(4 \max_j D_j)\right\rceil \cdot T_{\text{mix}}(\varepsilon) \cdot \prod_{j=1}^K (4D_j)$, where we set $T_{\text{mix}}(\varepsilon) := \max_j T^j_{\text{mix}}(\varepsilon)$.*

[5] For the sake of simplicity the algorithm was given for the case $\Psi_K = \varnothing$. It is obvious how to extend the algorithm when some parameters are known.

Proof. Let μ_j be the stationary distribution of arm j. It is well-known that the expected *first return time* $\tau_j(s)$ in state s satisfies $\mu_j(s) = 1/\tau_j(s)$. Set $\tau_j := \max_s \tau_j(s)$, and $\tau := \max_j \tau_j$. Then, $\tau_j \leq 2D_j$.

Now consider the following scheme to reach a given state $(s_j, n_j)_{j=1}^K$: First, order the states (s_j, n_j) descendingly with respect to n_j. Thus, assume that $n_{j_1} > n_{j_2} > \ldots > n_{j_K} = 1$. Take $T_{\text{mix}}(\varepsilon)$ samples from arm j_1. (Then each arm will be ε-close to the stationary distribution, and the probability of reaching the right state s_{j_i} when sampling arm j_i afterwards is at least $\mu_{j_i}(s_{j_i}) - \varepsilon$.) Then sample each arm $j_2, j_3, \ldots$ exactly $n_{j_{i-1}} - n_{j_i}$ times.

We first show the lemma for $\varepsilon \leq \mu_0 := \min_{j,s} \mu_j(s)/2$. As observed before, for each arm j_i the probability of reaching the right state s_{j_i} is at least $\mu_{j_i}(s_{j_i}) - \varepsilon \geq \mu_{j_i}(s_{j_i})/2$. Consequently, the expected number of restarts of the scheme necessary to reach a particular state $(s_j, n_j)_{j=1}^K$ is upper bounded by $\prod_{j=1}^K 2/\mu_j(s_j)$. As each trial takes at most $2T_{\text{mix}}(\varepsilon)$ steps, recalling that $1/\mu_j(s) = \tau_j(s) \leq 2D_j$ proves the bound for $\varepsilon \leq \mu_0$.

Now assume that $\varepsilon > \mu_0$. Since $D_\varepsilon \leq D_{\varepsilon'}$ for $\varepsilon > \varepsilon'$ we obtain a bound of $2T_{\text{mix}}(\varepsilon') \prod_{j=1}^K (4D_j)$ with $\varepsilon' := \mu_0 = 1/2\tau$. By (1), we have $T_{\text{mix}}(\varepsilon') \leq \lceil \log_2(1/\varepsilon') \rceil \, T_{\text{mix}}(1/4) \leq \lceil \log_2(4\tau) \rceil \, T_{\text{mix}}(\varepsilon)$, which proves the lemma. □

Proof of Theorem 2. Note that in each arm j the support of the transition probability distribution is upper bounded by $|S_j|$. Hence, Theorem 4 with $C_U = \sum_{j=1}^K |S_j| \, T^j_{\text{mix}}(\varepsilon)$ and $B = \max_j |S_j|$ shows that the regret is bounded by $42 D_\varepsilon \sqrt{\max_i |S_i| \cdot \sum_{j=1}^K |S_j| \cdot T^j_{\text{mix}}(\varepsilon) \cdot T \log\left(\frac{T}{\delta}\right)} + \varepsilon(D_\varepsilon + 2)T$ with probability $\geq 1 - \delta$. Since $\varepsilon = 1/\sqrt{T}$, this proves the first bound by Lemma 1 and recalling (1).

If the horizon T is not known, guessing T using the doubling trick (i.e., executing the algorithm for $T = 2^i$ with confidence parameter $\delta/2^i$ in rounds $i = 1, 2, \ldots$) achieves the bound given in Theorem 2 with worse constants.

Similarly, if T_{mix} is unknown, one can perform the algorithm in rounds $i = 1, 2, \ldots$ of length 2^i with confidence parameter $\delta/2^i$, choosing an increasing function $a(t)$ to guess an upper bound on T_{mix} at the beginning t of each round. This gives a bound of order $a(T)^{3/2}\sqrt{T}$ with a corresponding additive constant. In particular, choosing $a(t) = \log t$ the regret is bounded by $O\big(S \cdot \prod_{j=1}^K (4D_j) \cdot \max_i \log(D_i) \cdot \log^{7/2}(T/\delta) \cdot \sqrt{T}\big)$ with probability $\geq 1 - \delta$. □

Remark 4. Whereas it is not easy to obtain upper bounds on the mixing time in general, for *reversible* Markov chains T_{mix} can be linearly upper bounded by the diameter, cf. Lemma 15 in Chapter 4 of [15]. While it is possible to compute an upper bound on the diameter of a Markov chain from samples of the chain, we did not succeed in deriving any useful results on the quality of such bounds.

Remark 5. Periodic Markov chains do not converge to a stationary distribution. However taking into account the period of the arms, one can generalize our results to the periodic case. Considering in an m-periodic Markov chain the m-step transition probabilities given by the matrix P^m, one obtains m distinct aperiodic chains (depending on the initial state) each of which converges to a stationary distribution with respective mixing times. The maximum over these mixing times can be considered to be *the* mixing time of the chain.

Thus, instead of aggregating states (s_j, n_j), (s'_j, n'_j) with $n_j, n'_j \geq T^j_{\mathrm{mix}}(\varepsilon)$ as in the case of aperiodic chains, one aggregates them only if additionally $n_j \equiv n'_j \mod m_j$. If the periods m_j are not known to the learner, one can use the least common denominator of $1, 2, \ldots, |S_j|$ as period. Since by the prime number theorem the latter is exponential in $|S_j|$, the obtained results for periodic arms show worse dependence on the number of states. (Concerning the proof of Lemma 1 the sampling scheme has to be slightly adapted so that one samples in the right period when trying to reach a particular state.)

Proof of Theorem 3. Consider K arms all of which are deterministic cycles of length m and hence m-periodic. Then the learner faces m distinct learning problems with K arms, each of which can be made to force regret of order $\Omega(\sqrt{KT/m})$ in the T/m steps the learner deals with the problem [4]. Overall, this gives the claimed bound of $\Omega(\sqrt{mKT}) = \Omega(\sqrt{ST})$. Adding a sufficiently small probability (with respect to the horizon T) of staying in some state of each arm, one obtains the same bounds for aperiodic arms. □

8 Extensions and Outlook

Unknown state space. If (the size of) the state space of the individual arms is unknown, some additional exploration of each arm will sooner or later determine the state space. Thus, we may execute our algorithm on the known state space where between two episodes we sample each arm until all known states have been sampled at least once. The additional exploration is upper bounded by $O(\log T)$, as there are only $O(\log T)$ many episodes, and the time of each exploration phase can be bounded with known results. That is, the expected number of exploration steps needed until all states of an arm j have been observed is upper bounded by $D_j \log(3|S_j|)$ (cf. Theorem 11.2 of [10]), while the deviation from the expectation can be dealt with by Markov inequality or results from [16]. That way, one obtains bounds as in Theorem 2 for the case of unknown state space.

Improving the bounds. All parameters considered, there is still a large gap between the lower and the upper bound on the regret. As a first step, it would be interesting to find out whether the dependence on the diameter of the arms is necessary. Also, the current regret bounds do not make use of the interdependency of the transition probabilities in the Markov chains and treat n-step and n'-step transition probabilities independently. Finally, a related open question is how to obtain estimates and upper bounds on mixing times.

More general models. After considering bandits with i.i.d. and Markov arms, the next natural step is to consider more general time-series distributions. Generalizations are not straightforward: already for the case of Markov chains of order (or memory) 2 the MDP representation of the problem (Section 5) breaks down, and so the approach taken here cannot be easily extended. Stationary ergodic distributions are an interesting more general case, for which the first question is whether it is possible to obtain asymptotically sublinear regret.

Acknowledgments. This research was funded by the Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER (Contrat de Projets Etat Region CPER 2007-2013), ANR projects EXPLO-RA (ANR-08-COSI-004), Lampada (ANR-09-EMER-007) and CoAdapt, and by the European Community's FP7 Program under grant agreements n° 216886 (PASCAL2) and n° 270327 (CompLACS). The first author is currently funded by the Austrian Science Fund (FWF): J 3259-N13.

References

1. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Adv. in Appl. Math. **6** (1985) 4–22
2. Akyildiz, I.F., Lee, W.Y.L.W.Y., Vuran, M.C., Mohanty, S.: A survey on spectrum management in cognitive radio networks. IEEE Commun. Mag. **46**(4) (2008) 40–48
3. Anantharam, V., Varaiya, P., Walrand, J.: Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays, part II: Markovian rewards. IEEE Trans. Automat. Control **32**(11) (1987) 977–982
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multiarmed bandit problem. SIAM J. Comput. **32** (2002) 48–77
5. Audibert, J.Y., Bubeck, S.: Minimax policies for adversarial and stochastic bandits. In: colt2009. Proc. 22nd Annual Conf. on Learning Theory. (2009) 217–226
6. Jaksch, T., Ortner, R., Auer, P.: Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. **11** (2010) 1563–1600
7. Bartlett, P.L., Tewari, A.: REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In: Proc. 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009, AUAI Press (2009) 35–42
8. Tekin, C., Liu, M.: Adaptive learning of uncontrolled restless bandits with logarithmic regret. In: 49th Annual Allerton Conference, IEEE (2011) 983–990
9. Filippi, S., Cappe and, O., Garivier, A.: Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds. IEEE J. Sel. Topics Signal Process. **5**(1) (2011) 68–76
10. Levin, D.A., Peres, Y., Wilmer, E.L.: Markov chains and mixing times. American Mathematical Society (2006)
11. Gittins, J.C.: Bandit processes and dynamic allocation indices. J. R. Stat. Soc. Ser. B Stat. Methodol. **41**(2) (1979) 148–177
12. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multi-armed bandit problem. Mach. Learn. **47** (2002) 235–256
13. Whittle, P.: Restless bandits: Activity allocation in a changing world. J. Appl. Probab. **25** (1988) 287–298
14. Ortner, R.: Pseudometrics for state aggregation in average reward Markov decision processes. In: Proc. 18th International Conf. on Algorithmic Learning Theory, ALT 2007, Springer (2007) 373–387
15. Aldous, D.J., Fill, J.: Reversible Markov Chains and Random Walks on Graphs. (in preparation) http://www.stat.berkeley.edu/∼aldous/RWG/book.html.
16. Aldous, D.: Threshold limits for cover times. J. Theoret. Probab. **4** (1991) 197–211

A Proof of Theorem 4

Splitting into Episodes We follow the proof of Theorem 2 in [6]. First, as shown in Section 4.1 of [6], setting $\varDelta_k := \sum_{s,a} v_k(s,a)(\rho^* - r(s,a))$ with probability at least $1 - \frac{\delta}{12T^{5/4}}$ the regret after T steps can be upper bounded by

$$\sum_{k=1}^{m} \varDelta_k + \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)}\,. \tag{5}$$

Failing Confidence Intervals Concerning the regret with respect to the true MDP M being not contained in the set of plausible MDPs $\mathcal{M}_k$, we cannot use the same argument (that is, Lemma 17 in Appendix C.1) as in [6], since the random variables we consider for rewards and transition probabilities are independent, yet not identically distributed.

Instead, fix a state-action pair (s,a), let $S(s,a)$ be the set of states s' with $p(s'|s,a) > 0$ and recall that $\hat{r}(s,a)$ and $\hat{p}(\cdot|s,a)$ are the estimates for rewards and transition probabilities calculated from all samples of state-action pairs of the same color $c(s,a)$. Now assume that at step t there have been $n > 0$ samples of state-action pairs of color $c(s,a)$ and that in the i-th sample action a_i has been chosen in state s_i and a transition to state s'_i has been observed $(i = 1,\dots,n)$. Then

$$\begin{aligned}
&\left\| \hat{p}(\cdot|s,a) - \mathbb{E}[\hat{p}(\cdot|s,a)] \right\|_1 = \sum_{s' \in S(s,a)} \left| \hat{p}(s'|s,a) - \mathbb{E}[\hat{p}(s'|s,a)] \right| \\
&\leq \sup_{x \in \{0,1\}^{|S(s,a)|}} \sum_{s' \in S(s,a)} \Big(\hat{p}(s'|s,a) - \mathbb{E}[\hat{p}(s'|s,a)] \Big)\, x(s') \\
&= \sup_{x \in \{0,1\}^{|S(s,a)|}} \frac{1}{n} \sum_{i=1}^{n} \Big(x(\phi_{s_i,a_i,s,a}(s'_i)) - \sum_{s'} p(s'|s_i,a_i) \cdot x(\phi_{s_i,a_i,s,a}(s')) \Big)\,.
\end{aligned} \tag{6}$$

For fixed $x \in \{0,1\}^{|S(s,a)|}$, $X_i := x(\phi_{s_i,a_i,s,a}(s'_i)) - \sum_{s'} p(s'|s_i,a_i) \cdot x(\phi_{s_i,a_i,s,a}(s'))$ is a martingale difference sequence with $|X_i| \leq 2$, so that by Azuma-Hoeffding inequality (e.g., Lemma 10 in [6]), $\Pr\{\sum_{i=1}^{n} X_i \geq \theta\} \leq \exp(-\theta^2/8n)$ and in particular

$$\Pr\Big\{ \sum_{i=1}^{n} X_i \geq \sqrt{56Bn \log\left(\tfrac{4tC_U}{\delta}\right)} \Big\} \leq \left(\tfrac{\delta}{4tC_U}\right)^{7B} < \tfrac{\delta}{2^B 20 t^7 C_U}.$$

Recalling that by assumption $|S(s,a)| \leq B$, a union bound over all sequences $x \in \{0,1\}^{|S(s,a)|}$ then shows from (6) that

$$\Pr\Big\{ \left\| \hat{p}(\cdot|s,a) - \mathbb{E}[\hat{p}(\cdot|s,a)] \right\|_1 \geq \sqrt{\tfrac{56B}{n} \log\left(4C_U t/\delta\right)} \Big\} \leq \tfrac{\delta}{20t^7 C_U}. \tag{7}$$

Concerning the rewards, as in the proof of Lemma 17 in Appendix C.1 of [6] — but now using Hoeffding for independent and not necessarily identically distributed random variables — we have that

$$\Pr\Big\{ \left| \hat{r}(s,a) - \mathbb{E}[\hat{r}(s,a)] \right| \geq \sqrt{\tfrac{7}{2n} \log\left(2C_U t/\delta\right)} \Big\} \leq \tfrac{\delta}{60t^7 C_U}. \tag{8}$$

A union bound over all t possible values for n and all C_U colors of states in Ψ_U shows that the confidence intervals in (7) and (8) hold with probability at least $1-\frac{\delta}{15t^6}$ for the actual counts $N(c(s,a))$ and all state-action pairs (s,a). (Note that equations (7) and (8) are the same for state-action pairs of the same color.)

By linearity of expectation, $\mathbb{E}[\hat{r}(s,a)]$ can be written as $\frac{1}{n}\sum_{i=1}^{n} r(s_i,a_i)$ for the sampled state-action pairs (s_i,a_i). Since the (s_i,a_i) are assumed to have the same color $c(s,a)$, it holds that $|r(s_i,a_i)-r(s,a)|<\varepsilon$ and hence $|\mathbb{E}[\hat{r}(s,a)]-r(s,a)|<\varepsilon$. Similarly, $\big\|\mathbb{E}[\hat{p}(\cdot|s,a)]-p(\cdot|s,a)\big\|_1<\varepsilon$. Together with (7) and (8) this shows that with probability at least $1-\frac{\delta}{15t^6}$ for all state-action pairs (s,a)

$$\big|\hat{r}(s,a)-r(s,a)\big| < \varepsilon + \sqrt{\tfrac{7}{2n}\log\left(2C_U t/\delta\right)}, \tag{9}$$

$$\Big\|\hat{p}(\cdot|s,a)-p(\cdot|s,a)\Big\|_1 < \varepsilon + \sqrt{\tfrac{56B}{n}\log\left(4C_U t/\delta\right)}. \tag{10}$$

Thus, the true MDP is contained in the set of plausible MDPs $\mathcal{M}(t)$ at step t with probability at least $1-\frac{\delta}{15t^6}$, just as in Lemma 17 of [6]. The argument that

$$\textstyle\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} \;\leq\; \sqrt{T} \tag{11}$$

with probability at least $1-\frac{\delta}{12T^{5/4}}$ then can be taken without any changes from Section 4.2 of [6].

Episodes with $\boldsymbol{M\in\mathcal{M}_k}$ Now assuming that the true MDP M is in $\mathcal{M}_k$, we first reconsider extended value iteration. In Section 4.3.1 of [6] it is shown that for the state values $u_i(s)$ in the i-th iteration it holds that $\max_s u_i(s)-\min_s u_i(s)\leq D$, where D is the diameter of the MDP. Now we want to replace D with the diameter D_ε of the aggregated MDP. For this, first note that for any two states s,s' which are aggregated we have by definition of the aggregated MDP that $u_i(s)=u_i(s')$. As it takes at most D_ε steps on average to reach any aggregated state, repeating the argument of Section 4.3.1 of [6] shows that

$$\max_s u_i(s) - \min_s u_i(s) \leq D_\varepsilon. \tag{12}$$

Let $\tilde{\boldsymbol{P}}_k := \big(\tilde{p}_k(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$ be the transition matrix of $\tilde{\pi}_k$ on $\tilde{M}_k$, and $\boldsymbol{v}_k := \big(v_k\big(s,\tilde{\pi}_k(s)\big)\big)_s$ the row vector of visit counts in episode k for each state and the corresponding action chosen by $\tilde{\pi}_k$. Then as shown in Sect. 4.3.1 of [6][6]

$$\Delta_k \leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k-\boldsymbol{I}\big)\boldsymbol{w}_k + \sum_{s,a} v_k(s,a)\big(\tilde{r}_k(s,a)-r(s,a)\big),$$

where $\boldsymbol{w}_k$ is the normalized state value vector with $w_k(s):=u(s)-(\min_s u(s)-\max_s u(s))/2$, so that $\|\boldsymbol{w}_k\|\leq\frac{D_\varepsilon}{2}$. Now for $(s,a)\in\Psi_K$ we have $\tilde{r}_k(s,a)=r(s,a)$, while for $(s,a)\in\Psi_U$ the term $\tilde{r}_k(s,a)-r(s,a)\leq|\tilde{r}_k(s,a)-\hat{r}_k(s,a)|+|r(s,a)-$

[6] Here we neglect the error by value iteration explicitly considered in Sect. 4.3.1 of [6].

$\hat{r}_k(s,a)|$ is bounded according to (2) and (9), as we assume that $\tilde{M}_k, M \in \mathcal{M}_k$. Summarizing state-action pairs of the same color we get

$$\Delta_k \le \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2\sum_{c\in C(\Psi_U)} v_k(c)\cdot\Big(\varepsilon + \sqrt{\tfrac{7\log(2C_U t_k/\delta)}{2\max\{1,N_k(c)\}}}\Big),$$

where $C(\Psi_U)$ is the set of colors of state-action pairs in Ψ_U. Let T_k be the length of episode k. Then noting that $N'_k(c) := \max\{1, N_k(c)\} \le t_k \le T$ we get

$$\Delta_k \le \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2\varepsilon T_k + \sqrt{14\log\big(\tfrac{2C_U T}{\delta}\big)}\sum_{c\in C(\Psi_U)}\frac{v_k(c)}{\sqrt{N'_k(c)}}. \tag{13}$$

The True Transition Matrix Let $\boldsymbol{P}_k := \big(p(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$ be the transition matrix of $\tilde{\pi}_k$ in the true MDP M. We split

$$\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k = \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k + \boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k. \tag{14}$$

By assumption $\tilde{M}_k, M \in \mathcal{M}_k$, so that using (3) and (10) the first term in (14) can be bounded by (cf. Section 4.3.2 of [6])

$$\begin{aligned}
\boldsymbol{v}_k\ \big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k\ &\le\ \sum_{s,a} v_k\big(s,a\big)\cdot\big\|\tilde{p}_k(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \cdot \|\boldsymbol{w}_k\|_\infty\\
&\le 2\sum_{c\in C(\Psi_U)} v_k\big(c\big)\cdot\Big(\varepsilon + \sqrt{\tfrac{56B\log(4C_U T/\delta)}{N'_k(c)}}\Big)\cdot\tfrac{D_\varepsilon}{2}\\
&\le \varepsilon D_\varepsilon T_k + D_\varepsilon\sqrt{56B\log\big(\tfrac{2C_U T}{\delta}\big)}\sum_{c\in C(\Psi_U)}\frac{v_k(c)}{\sqrt{N'_k(c)}},
\end{aligned} \tag{15}$$

since — as for the rewards — the contribution of state-action pairs in Ψ_K is 0.

Concerning the second term in (14), as shown in Section 4.3.2 of [6] one has with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\sum_{k=1}^{m}\boldsymbol{v}_k\ (\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k\mathbb{1}_{M\in\mathcal{M}_k} \le D_\varepsilon\sqrt{\tfrac{5}{2}T\log\big(\tfrac{8T}{\delta}\big)} + D_\varepsilon\, C_U\log_2\big(\tfrac{8T}{C_U}\big), \tag{16}$$

where m is the number of episodes, and the bound $m \le C_U\log_2(8T/C_U)$ used to obtain (16) is derived analogously to Appendix C.2 of [6].

Summing over Episodes with $M \in \mathcal{M}_k$ To conclude, we sum (13) over all episodes with $M \in \mathcal{M}_k$, using (14), (15), and (16), which yields that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned}
\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_k} &\le D_\varepsilon\sqrt{\tfrac{5}{2}T\log\big(\tfrac{8T}{\delta}\big)} + D_\varepsilon\, C_U\log_2\big(\tfrac{8T}{C_U}\big) + \varepsilon(D_\varepsilon + 2)T\\
&+ \Big(D_\varepsilon\sqrt{56B\log\big(\tfrac{2C_U BT}{\delta}\big)} + \sqrt{14\log\big(\tfrac{2C_U T}{\delta}\big)}\Big)\sum_{k=1}^{m}\sum_{c\in C(\Psi_U)}\frac{v_k(c)}{\sqrt{N'_k(c)}}.
\end{aligned} \tag{17}$$

As in Sect. 4.3.3 and Appendix C.3 of [6], one obtains $\sum_{c\in C(\Psi_U)}\sum_k \frac{v_k(c)}{\sqrt{N'_k(c)}} \leq \left(\sqrt{2}+1\right)\sqrt{C_U T}$. Thus, evaluating (5) by summing Δ_k over all episodes, by (11) and (17) the regret is upper bounded with probability $\geq 1-\frac{\delta}{4T^{5/4}}$ by

$$\sum_{k=1}^{m}\Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} + \sum_{k=1}^{m}\Delta_k \mathbb{1}_{M\in\mathcal{M}_k} + \sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)}$$
$$\leq \sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)} + \sqrt{T} + D_\varepsilon\sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)} + D_\varepsilon\, C_U \log_2\left(\tfrac{8T}{C_U}\right)$$
$$+\varepsilon(D_\varepsilon+2)T + 3\left(\sqrt{2}+1\right)D_\varepsilon\sqrt{14BC_U T\log\left(\tfrac{2C_U BT}{\delta}\right)}\,.$$

Further simplifications as in Appendix C.4 of [6] finish the proof. □